ARMY RESEARCH LABORATORY

Crowd Sourcing Data Collection through Amazon Mechanical Turk

by Cynthia Pierce and Nicholas Fung

ARL-MR-0848 **September 2013**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-MR-0848 **September 2013**

Crowd Sourcing Data Collection through Amazon Mechanical Turk

Cynthia Pierce and Nicholas Fung
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2013	Final	September 2011

4. TITLE AND SUBTITLE
Crowd Sourcing Data Collection through Amazon Mechanical Turk

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Cynthia Pierce and Nicholas Fung

5d. PROJECT NUMBER
1FWDKA

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-A
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-MR-0848

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Crowdsourcing is an increasingly popular technique used to complete complex tasks or collect large amounts of data. This report documents the effort to employ crowdsourcing using the Mechanical Turk service hosted by Amazon. The task was to collect labeling data on several thousands of short videos clips as such labels would be perceived by a human. The approach proved to be viable, collecting large amounts of data in a relatively short time frame, but required specific considerations for the population of workers and impersonal medium through which data were collected.

15. SUBJECT TERMS
Crowdsource, Amazon, mechanical turk

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Nicholas Fung
Unclassified	Unclassified	Unclassified	UU	38	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-3101

Contents

List of Figures

List of Tables

Acknowledgments

We would like to acknowledge and thank Laurel Sadler for the creation of the graphical user interfaces (GUIs) that were used to collect internal data for the Recognition Panel and Round Table studies and Robert Winkler for the creation and maintenance of the database used for data storage and organization. We would also like to acknowledge and thank Andrew Thompson for providing expert analysis of the data. Finally, we would like to thank members of the U.S. Army Research Laboratory (ARL) and the Asset Control and Behaviors Branch for serving as subjects for the Recognition Panel and Round Table studies.

1. Introduction

Crowdsourcing is an increasing popular method through which a large amount of data can be obtained from a variety of sources. This method involves tapping into a large population of test subjects through the use of the internet. Typically, users are gathered and tasked with answering a survey or completing a small task in exchange for a reward.

This report chronicles the design and implementation of a crowdsourcing effort mean to provide labels for a large number of short videos. Each of these videos depicts the demonstration of a specific verb. The data gathering will provide ground truth data in the form of action labels for these videos as interpreted by human respondents. Amazon Mechanical Turk was identified as the service to utilize for the crowdsourcing study. In addition, this report covers efforts to minimize the effects of users putting forth insincere efforts or providing unsatisfactory responses.

2. Vignette Labeling Project

2.1 Overview

The goal of this project is to generate a large number of video vignettes meant to visually demonstrate specific verbs. The project called for 48 verbs to be demonstrated, each in 10 different exemplars. Each exemplar was filmed with 16 different setting variations, consisting of different backgrounds, camera angle, time of day, etc. This produces a total of 7680 vignettes. The vignettes are to be data points provided to system design teams as part of a larger research and development project.

This report documents efforts to produce ground-truth labeling of these videos. While the videos were filmed with the intention of demonstrating specific verbs, it was important to establish whether this intent was sufficiently conveyed. In addition, demonstrating specific verbs often involve different verbs. As an example, a demonstration of the verb “run” also includes a depiction of the verb “move.” A number of tests were established to provide comprehensive labels and attempt to validate the intended verb depictions. Also under consideration was interpretation of the videos by multiple sources. A video may display actions depicted as “run” to some people, while it may depict “walk” to others.

2.2 Videos

The vignettes are of short duration, mostly 8 to 20 s in duration. The brevity of each video means that the activities performed are lacking in overall context. That is by design—they are intended to be very focused on the actions and not on possible contexts in which such actions might occur.

The 7680 videos were divided evenly into Evaluation and Development sets. The two sets were to serve different purposes for the overall project, although both required accurate labeling.

3. Human Response Data: Recognition Task

3.1 Overview

Human response data were collected across four different performance tasks to comprehensively label the video vignettes: Recognition Task, Roundtable Task, Description Task, and Gap Filling Task. These data are intended to aid in the development and evaluation of visual analysis systems that perceive and comprehend motion in human terms in an unsupervised setting.

Of these four tasks, the Recognition Task accounts for the majority of data collected. Three separate studies were conducted for this task to contrast methods, control quality, and explore the means by which human data might be collected. This section documents the process and motivation for each recognition study.

The first recognition study consisted of a *Panel Study* using a simple detection protocol, in which participants were presented with vignettes and, for each vignette, asked "Do you see X?" where X was one of the 48 verbs. The purpose of this Panel Study was to establish a baseline for human performance on this protocol that could be used to assess the quality of data obtained using a similar protocol via crowdsourcing. Due to the large number of candidate probes and small number of participants, it was necessary to conduct this first study on a small but representative subset of the recognition corpus. This subset included 24 of the 48 verbs, but only one exemplar from each verb. All 16 variants from each exemplar were used, delegating one variant to each participant.

Next, the U.S. Army Research Laboratory (ARL) conducted a *Round Table Study*, in which participants observed vignettes in a self-regulated fashion, during which they performed an exhaustive search, indicating every verb (from the set of 48) perceived to be in a given vignette. By effectively retrieving 48 detection responses from a single probe, it was possible to cover the full set of 480 exemplars, using one randomly chosen variant per exemplar. Thus, the Round Table Study enabled broad response data coverage for the recognition corpus, to be used in place of the crowdsourced data, should this data be found invalid.

As previously suggested, the number of exemplar-variant-probe combinations (368,640) was intractable for an ARL Panel study. Thus, a *Crowdsourced Study* that mirrored the protocol used in the panel study was implemented to collect a complete set of response data for the entire year 1 recognition corpus.

In the following sections, each study is described in detail, including population, protocol, and methods.

3.2 Recognition Task: Panel Study

3.2.1 Rationale

The Recognition Task: Panel Study (RT:PS) was conducted to generate high quality response data from a tractable portion of the recognition corpus, which could then be used to validate data from the Crowdsourced Study in anticipation of using crowdsourcing in future program years as the primary means of data collection.

3.2.2 Participants

The ground-truth effort conducted at ARL consisted of a panel of engineers that performed similar tasks that would be performed in the crowdsourced effort. Because ARL employees are not malicious subjects, these responses could be considered as "best effort" responses and would not require screening for malicious data. The panel makeup was composed of 16 engineers. Herein all members of the panel are referred as the ARL panel.

3.2.3 Stimuli

The vignettes covered by the ARL panel for the RT:PS were a subset of the full 48 verb-10 exemplar-16 variant set. Due to the scale of the corpus and number of concepts to be detected, it was not feasible to perform an exhaustive data gathering effort across all verb-exemplar-variant vignettes. Instead, a sample of the year 1 vignette set was selected to be reviewed by the ARL panel. These data would be valuable in that they could be used as a check against the crowdsourced data, as malicious or incompetent data obtained from crowdsourcing was a concern. The subset of exemplars used for the ARL panel consisted of 24 verbs, 1 exemplar per verb, and 16 variants per exemplar, for a total of 384 vignettes. The specific exemplars were chosen based on the initial video file release, C-D1a.

The exemplars chosen from this subset of the development corpus were selected based on release status, coverage across the verb sets, and verb action depiction. In general, these exemplars were chosen to minimize noise without introducing bias. The exemplars selected were bounce4, bury2, carry5, catch1, chase9, close6, collide2, dig2, enter2, flee3, fly10, follow1, give2, hit3, jump10, kick5, lift7, open4, pickup2, push2, raise9, run4, throw3, and walk9. Two of these exemplars were released as "half sets." The "half sets" had only 8 variations (of 16) released from the same exemplar. These half sets were released to enable an evaluation of within-exemplar learning and performance during the year 1 evaluations.

3.2.4 Method

The ARL panel was conducted locally at the Adelphi Laboratory Center in Adelphi, MD. Software was developed and employed to present videos to ARL panel members and retrieve their response data. The graphical user interface (GUI) for the RT:PS presented a single verb question with a determination of present or absent for the action occurring in the vignettes. This protocol was known as Single Verb Present/Absent (SVPA).

For the RT:PS, each ARL panel member was shown a vignette and asked “Do you see X?” where X was one of the 48 verbs of interest, to ascertain if they observed a particular action occurring in the video. A definition of the action X was displayed at the same time to provide guidance as to limit the ambiguity of verbs that may have different interpretations. The panelist provided a “yes” or “no” response, corresponding to a judgment that the activity was present or absent.

The assignment of the vignettes to particular panel members is illustrated in table 1. The assignments were made so that each of the 16 variants was seen by one of the panel members. The column headers indicate the verb question that was asked for a particular variant. The P# indicates the identification number assigned to a panel member. For example, of the exemplar chosen to represent “approach,” the first variant was assigned to panelist P1, the second variant was assigned to panelist P2, etc.

Table 1. Vignette assignment.

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	…48
Verb	Exemplar	Variant	approach	arrive	attach	bounce	bury	carry	catch	chase	close	collide	dig	drop	enter	exchange	exit	fall	…walk
VERB (i)	Exemplar (j)	Var 1	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	…P16
		Var 2	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P1	…P1
		Var 3	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P1	P2	…P2
		Var 4	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P1	P2	P3	…P3
		Var 5	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P1	P2	P3	P4	…P4
		Var 6	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P1	P2	P3	P4	P5	…P5
		Var 7	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P1	P2	P3	P4	P5	P6	…P6
		Var 8	P8	P9	P10	P11	P12	P13	P14	P15	P16	P1	P2	P3	P4	P5	P6	P7	…P7
		Var 9	P9	P10	P11	P12	P13	P14	P15	P16	P1	P2	P3	P4	P5	P6	P7	P8	…P8
		Var 10	P10	P11	P12	P13	P14	P15	P16	P1	P2	P3	P4	P5	P6	P7	P8	P9	…P9
		Var 11	P11	P12	P13	P14	P15	P16	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	…P10
		Var 12	P12	P13	P14	P15	P16	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	…P11
		Var 13	P13	P14	P15	P16	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	…P12
		Var 14	P14	P15	P16	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	…P13
		Var 15	P15	P16	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	…P14
		Var 16	P16	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	…P15

Each panel member was assigned a set of vignettes to review based on table 1. The panel member reviewed the assigned vignette three times, once for three different verb questions. Once the assignments were made for each panel member, the vignettes were randomly presented in a GUI shown in figure 1.

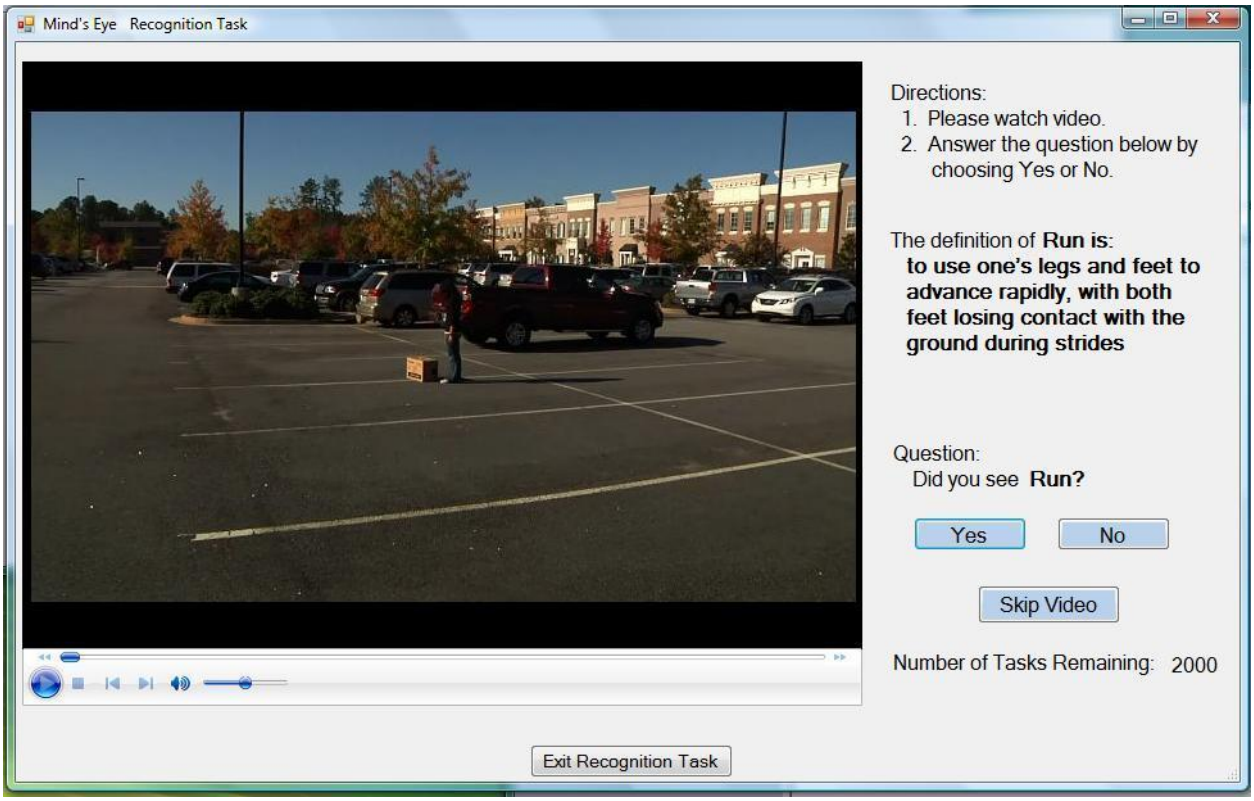


Figure 1. Panel study GUI.

The panel members were provided an identical set of instructions on each page of the GUI. In addition, they were all briefed of the project and the overall goal of the RT:PS. No additional training was provided. The definition of the verb asked was presented in the GUI in order to ensure consistent interpretation of the task across the user population. Each panel member was then asked a single verb question with a present or absent response. When a panel member selected their response from the GUI, the present or absent judgment was recorded in a central database and a new vignette with a different verb question would be immediately presented to the panel member. ARL also implemented a "Skip Video" button in the GUI, enabling the member to record no response to that particular question and move on to the next. Skipped stimuli vignettes would be replayed for that member at a later time until all required data were collected.

3.2.5 Resulting Data

The data collected from the RT:PS were stored in a central database and written to a human response (HR) file in compliance with a Broad Evaluation Plan (BEP) issued as part of the larger

project. Every response presented in the HR file contains the Stimulus ID, the Human ID, the verb question asked and the present/absent judgment recorded as a Boolean, where 1 represents "present."

The total number of responses for the ARL panel data was 17,664 responses. Twenty-four exemplars (two of which were half sets as described above) were covered through the ARL panel study. There was one response per variant per verb question asked with a single present/absent response. The calculation was as follows: 48 verb questions were asked regarding 16 variants of 22 exemplars, and 8 variants of 2 exemplars, and each vignette was viewed by only 1 reviewer, resulting in 48 x ((22x16) + (2x8)) = 17,664 total responses. The 16 ARL panel members each viewed 1,104 vignettes.

All of the data from the ARL panel were distributed to system development teams as part of a larger Development Set of data.

3.2.6 Assessment

The objective of this RT:PS, was to establish a reliable human performance baseline for the recognition set with respect to the 48 verb classes.

In order to assess the quality of the RT:PS data, we examined the distribution of "Present" responses (in the "Present"/"Absent" response dichotomy) over all 16 participants. Two outliers were identified on both extremes, one high (303) and one low (91), as indicated by the two red crosses in the box plot found in figure 2.

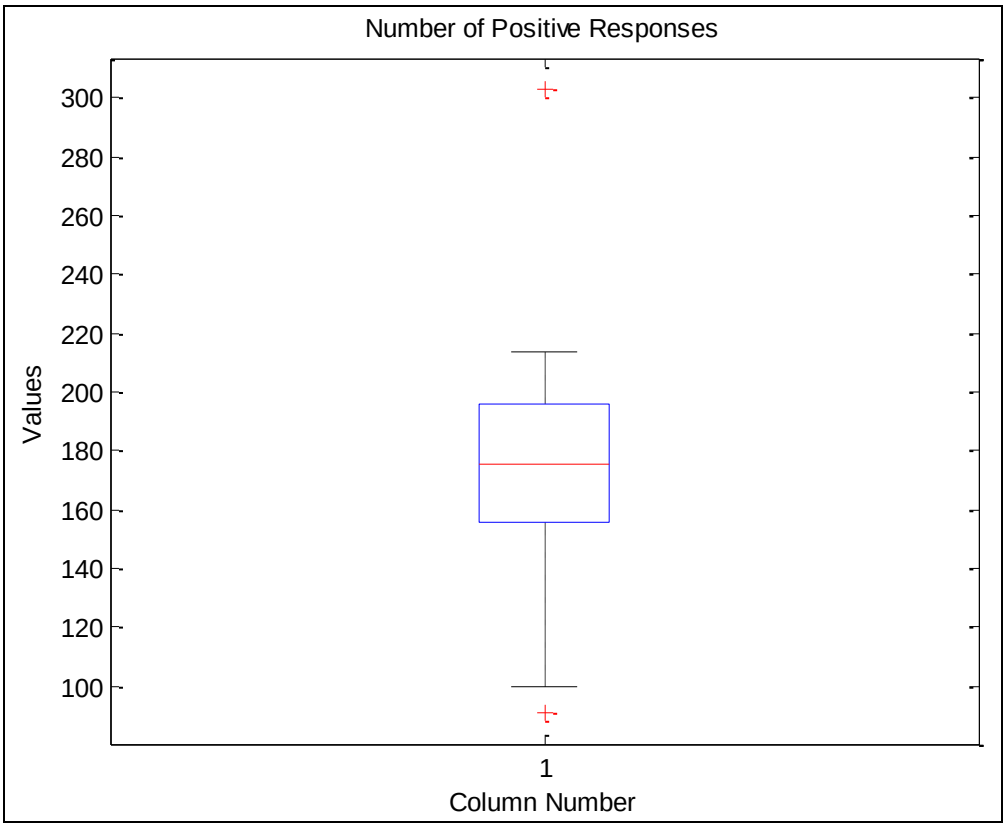


Figure 2. Outliers in recognition panel responses.

A closer examination of these outliers was undertaken to verify that participation in both cases was valid or to identify the data as erroneous. Follow-up interviews with the participants revealed only differences in response bias. The high response outlier approached the task significantly more sensitive to detections than the other panelists, reporting even minor detections of the requested action. The low response outlier focused on what was perceived as the main action in the video and chose not to report actions that were deemed incidental to the intent of the vignette.

The 48 verbs used to motivate stimulus vignette production, were expected to serve as semantically relevant dimensions of variance among the stimuli. There was no expectation that these dimensions are independent or complete. However, they may be sufficiently broad in scope to adequately define a similarity space that, if emulated by machines, could give rise to useful and general activity discrimination.

A preliminary look at the panel data begins to characterize the richness of this space. As described earlier, stimuli for this study were drawn from only 24 stimulus production classes. A principal components analysis of the RT:PS data reveals that 12 principal components account for 93% of the variation in the data. This suggests both that the stimuli are rich and varied and that the dimensions over which they are being assessed are broadly useful in their discrimination.

3.3 Recognition Task: Round Table Study

The Recognition Task: Round Table (RT:RT) study was conducted to collect a broader data set than was achieved by the RT:PS. Unlike the Panel Study, the Round Table members saw samples of exemplars from both the Development and Evaluation sets. The RT:RT study thus enabled broad response data coverage for the recognition corpus. The data could also be used as a sample to compare to the crowdsourced data to be collected later.

3.3.1 Rationale

The rationale for the RT:RT data collection effort was to gather data across all 48 verbs, each with 10 exemplars and one variant, randomly chosen from 16. All of the exemplars examined by the previous RT:PS were drawn from the Development set. The RT:RT effort provided data on vignettes from both the Development and Evaluation data sets. While these data were more sparsely collected across fewer exemplars, they provided data points more representative of the entire vignette set for comparison against future crowdsourced data.

3.3.2 Participants

The RT:RT members were a subset of five members from the RT:PS. All the RT members were familiar with the goals of the program and thus were considered trusted reviewers. Because the RT:RT was conducted after the RT:PS, it was expect that these members would be more familiar with the experiment. It was emphasized to these reviewers that their best effort must be put forth for this data collection because it could potentially be used as an evaluation set.

3.3.3 Stimuli

The vignettes covered by the RT:RT were a subset of the full 48 verb-10 exemplar-16 variant set. Given the desire to cover exemplars in both the Development and Evaluation sets and the need to create a manageable set of vignettes given time and resource constraints, it was decided that the panel would review one variant from each of the 480 verb-exemplars. These variants were selected randomly from the 16 variants associated with each verb-exemplar. Every member of the panel saw the same 480 videos, presented in a random order.

3.3.4 Method

The RT:RT members were asked to watch a video vignette and then indicate every verb present in the video. The videos were presented using a similar GUI to that of the RT:PS, and can be seen in figure 3. Each video could be replayed as many times as the panel member required. The cognitive load on the ARL RT:RT member was significantly higher under this protocol, as each subject was required to review all of the 48 possible verbs, as opposed to the Recognition Panel Study in which the respondent was presented with a single present/absent verb question. Each RT:RT participant took part in the task independently and the results from all participants were compared after the data collection was complete.

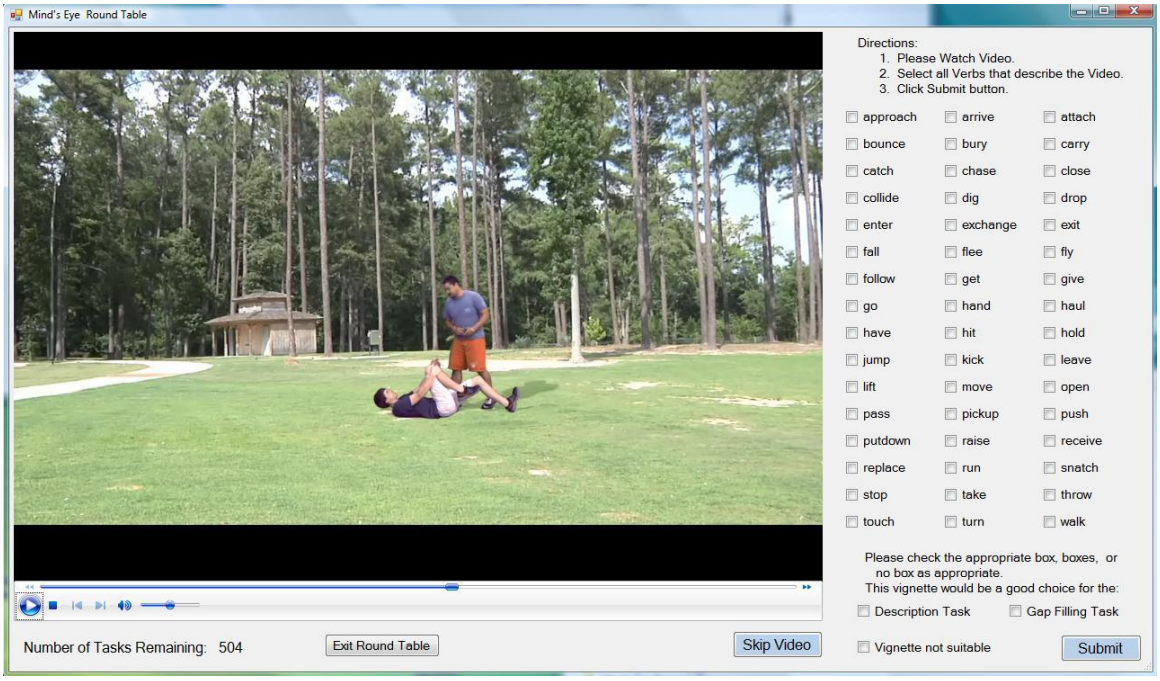


Figure 3. RT study GUI.

After selecting all applicable verbs for a given vignette, the RT:RT members would then record their responses to the database by pressing the Submit button. Once the RT member had submitted a response, the next vignette in the queue would be presented.

The RT:RT task was considered by the subjects to be difficult to perform when compared to the RT:PS because the RT:RT task required a significant amount of effort in that each member must consider all 48 verb definitions for each video displayed. While the verb definitions were provided, the subjects reported difficulty in remembering verb definitions. This feedback led to changes in the interfaces for the remaining Description and Gap-Filling tasks. For these tasks, the definitions would be displayed when the mouse passed over a verb in the GUI.

3.3.5 Resulting Data

The data collected from the RT:RT were stored in the database and recorded in separate HR files. Each HR file contained the Stimulus ID, the Human ID, the verb code, and the present/absent judgment as a Boolean for a set of responses. The format is identical to the SVPA Recognition data.

The task produced 115,200 data points (48 verbs, 10 exemplars, 1 variant, 48 verbs selections viewed by 5 ARL reviewers). The RT:RT provided data across all 480 exemplars, while the RT:PS only covered 24 exemplars. The 24 exemplars were selected from the development vignette set, producing a total of 58,080 data points. The remaining 57,120 responses were collected from the evaluation vignette set.

3.3.6 Assessment

The RT:RT Study was conducted to mitigate perceived risks associated with data collection and data validation in the Crowdsourcing Study. In the event that valid, accurate data could not be obtained from the crowdsourcing effort, the RT:RT data were intended to serve as training and evaluation data for the year 1 vignette set. As reported below, the crowdsourced data were later found to be valid for this evaluation. For this reason, the RT:RT Study data have not yet been formally evaluated.

Because a different response protocol was employed, it will be of interest to see how the data from this study compare to the data from other Recognition Studies. In particular, during a given stimulus presentation, participants in the RT Study listed all of the 48 candidate verbs. This response protocol results in the acquisition of data more quickly than the RT:PS or Crowdsourced Study protocol. If the RT:RT protocol produces data of similar quality to that of the other recognition studies, it may be a promising candidate for future studies.

3.4 Recognition Task: Crowdsourced Study

The Recognition Task: Crowdsourced Study (RT:CS) was the most comprehensive data collection effort undertaken. All variants for all exemplars across all production verb categories were included in this study. With 7676 video vignettes (4 vignettes were mistakenly not produced) and 48 verbs, there are a total of 368,448 different data points to be collected. In order to establish a data set that would allow for sufficient statistical power in a factorial analysis, it was necessary to employ a much larger workforce of the sort accessible through crowdsourcing resources. Amazon Mechanical Turk (forthwith referred to as “MTurk”) was identified as an appropriate medium for this data gathering task.

MTurk is an online service provided by Amazon. The service allows a Requester to form a typically small or easy task, called Human Intelligence Tasks (HITs), and solicit responses from Workers. Workers perform the task and, if completed satisfactorily, are compensated monetarily by the Requester. This exchange allows for Requesters to tap a large resource of human input in an efficient and cost-effective manner.

3.4.1 Rationale

Crowdsourcing was determined to be the most cost-effective and time-efficient method for collecting data over the vignette views needed. This study required 368,448 responses to be collected. Gathering these data would have been impossible with the small ARL population in the available time frame. Such an undertaking would have taken approximately 20 weeks of full-time panel effort. In contrast, it took approximately 10 days via crowdsourcing to collect this data, for a fraction of the cost. In addition, the ARL population was homogeneous as it consisted solely of engineers. Crowdsourcing offered access to a more diverse population. ARL workers were also potentially biased, as they have knowledge of the project including the motivation and long-term goals.

3.4.2 Participants

MTurk was chosen because it affords an easy mechanism to employ a large, diverse, on-demand workforce. However, there are potential issues involved with the participants likely to respond on MTurk. Participation was limited to MTurk Workers residing in the United States and who had a 95% approval rating. Because the task involves labeling the vignettes with English language verbs, the workers had to be familiar with the English language. For example, non-native speakers may not understand the distinction between "pull" and "haul." By limiting Workers to the United States, the goal was to accept only Workers that regularly use conversational level English skills. Participation was also limited to MTurk Workers whose historical approval rating was at least 95%. An individual's approval rating is highly valued by MTurk Workers, so they are conscientious about maintaining that rating. Workers need to maintain a high approval rating so that they can continue to be granted access to as many MTurk HITs as possible. For this reason, a high minimum approval rating will only elicit responses from high quality Workers seeking to perform at a high level. No identifying information was maintained about the MTurk Workers except for the numerical user IDs assigned for data storage in the ARL database and the WorkerID provided by MTurk.

3.4.3 Stimuli

All of the vignettes from both the Development and Evaluation sets were used in the crowdsourcing study. Thus, the full 480-exemplar, each with 16 variant set of vignettes was presented, each paired with 48 instances of the "Do you see X?" question, where "X" is one of the 48 verbs. This results in a total of 368,640 stimuli. A minor deviation to this number is discussed in section 2.4.5.

3.4.4 Method

The RT:CS was modeled after the RT:PS. MTurk Workers did not receive any training prior to attempting the task. For each task, also known as a stimulus, a vignette was displayed along with a verb question ("Do you see [verb]?") and the verb definition. Workers responded to a single verb question with a present/absent judgment just as in the RT:PS. In order to facilitate Worker productivity, each HIT displayed 20 different stimuli. This format limits the Worker overhead of having to find and open each HIT. One embellishment unique to the Crowdsourced protocol is the designation of a portion of the stimuli (10%) to serve as quality control checks. The method for this is described below.

3.4.4.1 Mechanical Turk HITs

In MTurk vernacular, a MTurk HIT is the smallest unit of work that results in a payment reward. HITs in general can be of varying length and offer varying rewards. In addition, each HIT has a given number of assignments. A given Worker typically only responds to one assignment per HIT; however, a single Worker may respond to multiple HITs from the same requester.

After completing the HIT, the Worker can respond to other HITs from the same Requester. It is important to make HITs appealing to Workers so that they are motivated to perform well and take on additional HITs, minimizing the overhead involved with a Worker becoming familiar with the task and interface. There are a number of ways to make HITs more appealing, such as high-reward payments, interesting tasks, and prompt approval systems.

3.4.4.2 Recognition HIT

For the purposes of this data collection task, HITs were composed of 20 stimulus presentations, each displaying a single vignette and requiring a single response to a present/absent query for a single verb. After a short set of instructions, the Worker observed each vignette, answering the associated present/absent question in turn, until all 20 questions had been answered. At this point, the Worker could revise each response or submit the HIT as complete. A portion of a sample HIT can be seen in figure 4.

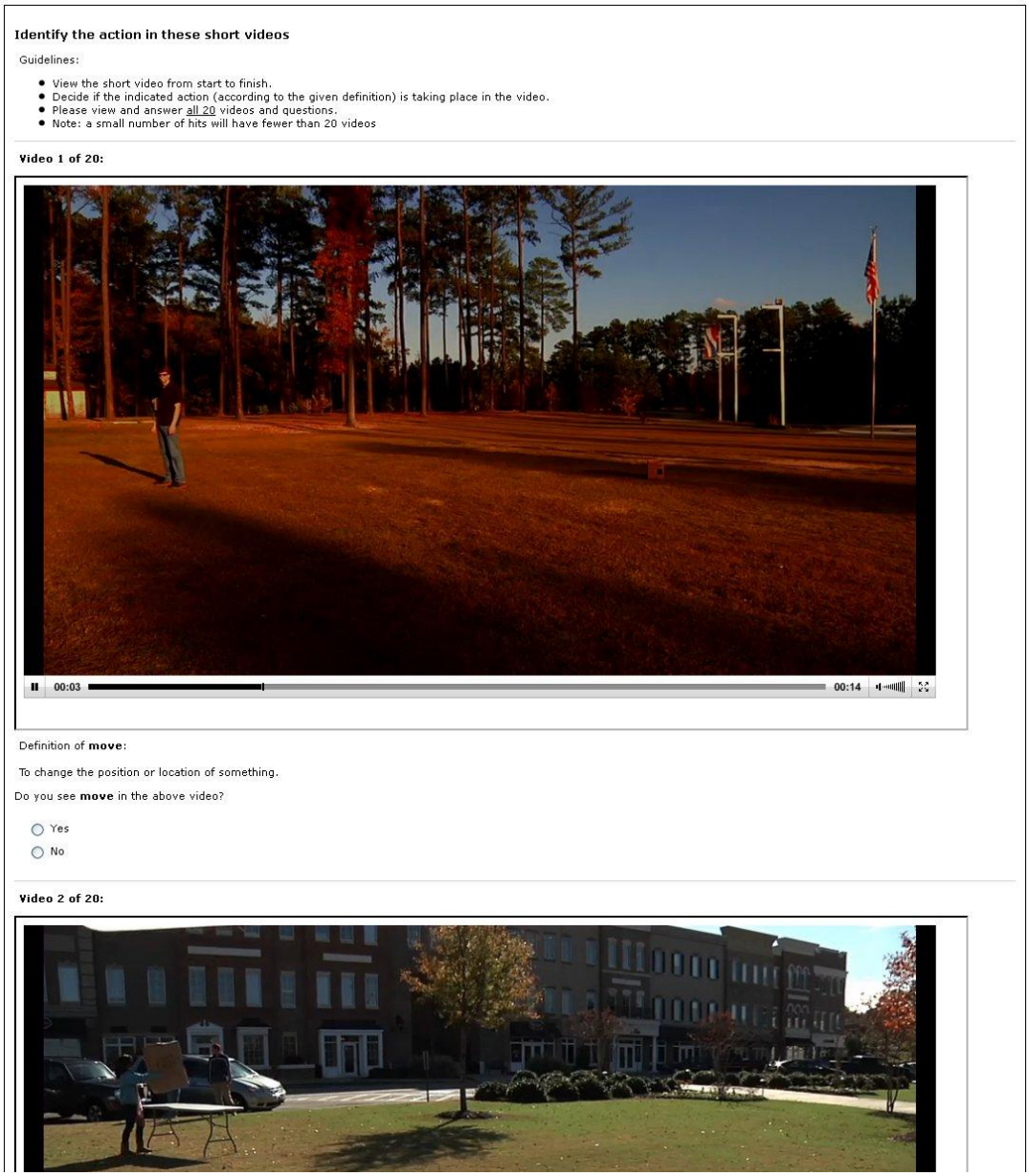


Figure 4. MTurk recognition task GUI.

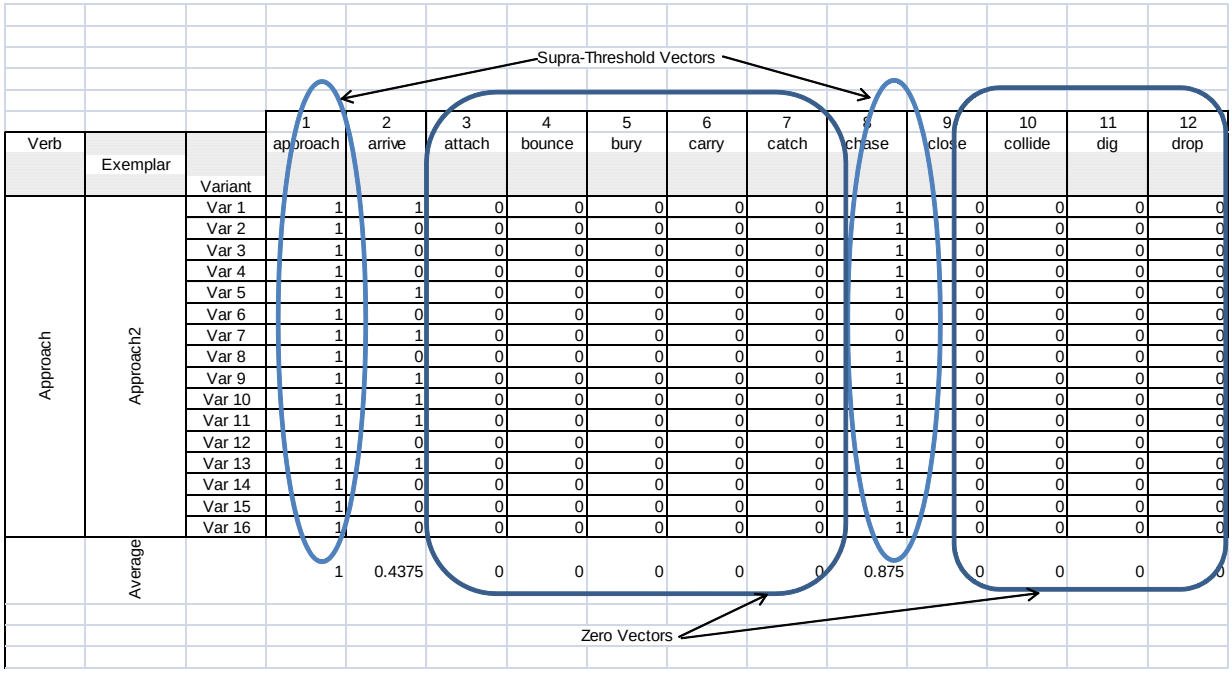
3.4.4.3 Check Vignettes

One danger of employing a crowdsourcing service such as MTurk is the danger of poor or insincere test results. The primary motivation of a Worker is, presumably, to obtain a maximum amount of reward payment. For this reason, the speed with which they complete HITs can take priority over accuracy. In addition, malicious workers may attempt to cheat the system by skipping steps or guessing rather than making genuine effort.

In order to provide a means for both real-time and post-collection quality control, a system of check stimuli was employed. These check stimuli were selected to have an expected positive or

negative response. Each HIT of 20 stimuli contained 2 check stimuli (18 data gathering stimuli). The Worker was not told that check vignettes were included in each HIT and had no indication which stimuli were check stimuli.

The check stimuli were selected based on data gathered from the ARL panel task. Recall that in the RT:PS, for a given exemplar, a data point was collected for each of the 16 variants over all of the 48 verbs for a total of 768 data points. The 16 responses collected for a given verb exemplar (one response per variant) were averaged for each exemplar-verb question pair (figure 5). If the mean value of the vector was zero, that was interpreted as agreement that the verb in question was not present in the exemplar vignettes, and all the vignettes in the vector could be used as false positive check vignettes. If the mean value of the vector was greater than 0.80, the vector was considered a Supra-Threshold vector. All the vignettes in the Supra-Threshold vector could be used as false negative check vignettes.



Verb	Exemplar	Variant	1 approach	2 arrive	3 attach	4 bounce	5 bury	6 carry	7 catch	8 chase	9 close	10 collide	11 dig	12 drop
Approach	Approach2	Var 1	1	1	0	0	0	0	0	1	0	0	0	0
		Var 2	1	0	0	0	0	0	0	1	0	0	0	0
		Var 3	1	0	0	0	0	0	0	1	0	0	0	0
		Var 4	1	0	0	0	0	0	0	1	0	0	0	0
		Var 5	1	1	0	0	0	0	0	1	0	0	0	0
		Var 6	1	0	0	0	0	0	0	0	0	0	0	0
		Var 7	1	1	0	0	0	0	0	0	0	0	0	0
		Var 8	1	0	0	0	0	0	0	1	0	0	0	0
		Var 9	1	1	0	0	0	0	0	1	0	0	0	0
		Var 10	1	1	0	0	0	0	0	1	0	0	0	0
		Var 11	1	1	0	0	0	0	0	1	0	0	0	0
		Var 12	1	0	0	0	0	0	0	1	0	0	0	0
		Var 13	1	1	0	0	0	0	0	1	0	0	0	0
		Var 14	1	0	0	0	0	0	0	1	0	0	0	0
		Var 15	1	0	0	0	0	0	0	1	0	0	0	0
		Var 16	1	0	0	0	0	0	0	1	0	0	0	0
	Average		1	0.4375	0	0	0	0	0	0.875	0	0	0	0

Figure 5. Threshold vectors.

When an MTurk Worker attempted a HIT, there were three possible outcomes with respect to these check vignettes: the Worker could get both of the check vignettes correct, the Worker could answer one but not both correct, or neither check stimuli question could be answered correctly. Our payment policy was that if one of the two check stimuli was answered incorrectly, the MTurk Worker was paid but the HIT was resubmitted for new data. If both check stimuli were answered incorrectly, the MTurk Worker was not paid and the HIT was resubmitted for new data. Workers whose HITs had both check vignettes correct were paid, their data were recorded, and the HIT was not resubmitted for additional data collection.

3.4.4.4 HIT Creation

The HITs were created using the MTurk C# application programming interface (API). MTurk defines HITs, including the number of assignments and how much a Worker can earn by responding in addition to account information, through extensible markup language (XML) code. Because XML strings can be formed and transmitted to MTurk within C# code using the MTurk libraries, forming a large number of HITs is possible within a single program.

The technical aspects of the program largely involve interacting with the database and forming the question XML strings. Because each HIT is identical in structure, varying only in the specific videos and question verbs, a template was created to be used for each HIT XML string. The template, depicted in figure 6, employs keyword tags such as "[url]" in place of the question specific string such as the exact URL for a specific video file. The program accomplishes this task through a query to the database to obtain video, verb pairing information, and replacement of the tags in the XML template through a simple substring replacement call using the C# string library. Multiple question XML strings can then be concatenated to form a HIT with more than one question. Once verified, the XML strings were uploaded to the MTurk service using the MTurk C# libraries over an HTTPS connection. These HITs are immediately published and made available for Workers to accept and complete.

```
<Question>
 <QuestionIdentifier>Q_[identifier]</QuestionIdentifier>
 <IsRequired>true</IsRequired>
 <QuestionContent>
  <FormattedContent>
   <![CDATA[
    <p>Definition of <b>[verb]</b>:<br></br>
    [definition]</p>
    Do you see <b>[verb]</b> in the above video?
   ]]>
  </FormattedContent>
 </QuestionContent>
 <AnswerSpecification>
  <SelectionAnswer>
   <StyleSuggestion>radiobutton</StyleSuggestion>
   <Selections>
    <Selection>
     <SelectionIdentifier>A_[identifier]_Y</SelectionIdentifier>
     <Text>Yes</Text>
    </Selection>
    <Selection>
     <SelectionIdentifier>A_[identifier]_N</SelectionIdentifier>
     <Text>No</Text>
    </Selection>
   </Selections>
  </SelectionAnswer>
 </AnswerSpecification>
</Question>
```

Figure 6. HIT XML string template.

The MTurk XML schema allows for blocks of extensible hypertext markup language (XHTML) code within the FormattedContent XML tags. This XHTML is similar to hypertext markup language (HTML) used in typical web development; however, the FormattedContent tag supports only a limited subset of XHTML. Notably, it does not support JavaScript. However, JavaScript (figure 7) was needed to display the vignettes to the workers. Instead, an inline frame (iframe) was used to load an external HTML file hosted on the Amazon Simple Storage Service (S3).

```
<Overview>
 <FormattedContent>
  <![CDATA[
      <p><b>[Video]:</b></p>
      <iframe src="[url]video.html?video='[movie]'"
      width="1000" height="580">
      If you can see this, your browser does not support
      IFRAME and you should not respond to this HIT!
      </iframe><br />
   ]]>
 </FormattedContent>
</Overview>
```

Figure 7. The iframe code.

This external HTML page is able to use JavaScript because it is a separate page that does not need to parse according to the MTurk XML code. Instead, this page is free to use JavaScript to embed JW Player, a Flash-based streaming video player. In order to load the appropriate vignette, the iframe code in the XML passes a variable to the external HTML file via a url variable. In the code block in figure 7, the iframe source page contains two tags, “[url]” and “[movie]”. The [url] tag is replaced with the root Amazon S3 location where the video loading HTML page is located, and the [movie] tag is the specific filename of the vignette that is to be loaded. The HTML file can parse the filename from the url using JavaScript string manipulation (figure 8).

The code block in figure 8 is the “video.html” file located on the Amazon S3 server. The JavaScript first parses the url string and assigns appropriate variable values. The script then embeds the JW Player using the variable values. JW Player is a streaming media library that can be embedded into a Web page. It creates a simple user interface including play/pause button and movie timeline. This interface is similar to that used by YouTube, providing a familiar interface for the Workers.

```
<SCRIPT LANGUAGE="JavaScript">
urlstring = window.location.search
urlstring = unescape(urlstring)
videoIndex = urlstring.indexOf("video")
video =
urlstring.substring(videoIndex+7,urlstring.indexOf("'",videoIndex+7))
video = "http://s3.amazonaws.com/ARL_ME/videos/"+video

document.write("<object classid=\"clsid:D27CDB6E-AE6D-11cf-96B8-
444553540000\" width=\"960\" height=\"540\" id=\"player1\"
name=\"player\"> <param name=\"movie\" value=\"jwplayer/player.swf\">
<param name=\"allowfullscreen\" value=\"true\"> <param
name=\"allowscriptaccess\" value=\"always\"> <param
name=\"flashvars\" value=\"file="+video+"&autostart=false\"> <embed
id=\"player1\" name=\"player1\" src=\"jwplayer/player.swf\"
width=\"960\" height=\"540\" allowscriptaccess=\"always\"
allowfullscreen=\"true\"
flashvars=\"file="+video+"&autostart=false\"/> </object><br></br>")
</SCRIPT>
```

Figure 8. The "video.html" file located on the Amazon S3 server.

3.4.4.5 Obtaining HIT Results

Obtaining the results, verifying the integrity, and rewarding the Workers follow a similar process to HIT creation. Upon creation of the HIT, MTurk generates a unique ID key. A query containing a HIT ID key can be made to MTurk to obtain all results associated with that HIT. Each result is stored in an XML string that must be parsed using the C# XML library tools according to the XML schema provided in the MTurk libraries. The data can then be sorted and stored in the local database.

One concern for collecting data was making the HITS attractive to prospective MTurk Workers. There were three factors that the MTurk Workers were concerned about: appropriate payment for effort, maintaining their 95% accuracy rating, and quick payment. The actual HIT cost 50 cents to view 20 vignettes. We were prepared to pay up to 80 cents, but that was not necessary as the rate of completion moved along quickly at 50 cents. We paid immediately those MTurk Workers that had both check vignettes correct. For the MTurk Workers that had one check vignette correct, the payment was delayed by 2 weeks. For those MTurk Workers that got both check vignettes incorrect, they were not paid. Because we picked a population that had a 95% accuracy rating, maintaining that rating was a concern for the MTurk Workers. Keeping those considerations in mind, we did not reject the MTurk Workers that had one check vignette wrong for fear of ruining their rating, which would, in turn, make them hesitant to complete additional HITS. Initially, MTurk Workers would only do a few HITs and see how quickly payment was made. As they had more confidence in meeting the requirements to get paid, the rate of HIT completion increased. In fact, the HITs were so popular the MTurk Workers completed all of the available HITs and requested that new HITS be posted.

After looking at the initial check vignette data, it seemed that about 83% of the MTurk population correctly answered both check vignette questions, 16.5% got were correct on one check vignette question, and only 0.5% answered neither check vignette questions correct. A later analysis revealed that some of the MTurk Workers had contributed spurious data (about 8% of the data)—when we analyzed all of the data contributed, it looked as if they were guessing, even for HITs where both check vignettes were answered correctly. All data from such workers were discarded and the HITs were resubmitted on MTurk. The methods used to determine this spurious data are explained in section 2.4.6.

3.4.5 Resulting Data

The entire year 1 corpus of 7,676 action vignettes, composed of 480 exemplars with 16 variants each, was used in the crowd sourced data collection effort. For each of these 7,676 vignettes, 48 present/absent judgments were collected corresponding to the 48 actions in the study set. Thus, a total of 368,640 (480 exemplars x 16 variants x 48 verb questions) responses would have been collected. However, one of the 480 exemplars ("FLY4") was incomplete by four vignettes. Therefore, the actual number of responses collected was 368,448.

The corpus has been divided into two parts: a Development Set and an Evaluation Set. The Development Set consists of slightly less than half of the original corpus, with approximately 4.5 exemplars (out of ten) per production verb class. Thus, the total number of recognition responses corresponding to the Development Human Response Set is 167,040. These data were be used in support of research and system development. The remaining 201,408 responses constitute the Evaluation Human Response Set and are used in to compare system performance with human performance.

3.4.6 Assessment

3.4.6.1 Data Cleansing

The initial assessment goal was to identify and remove sources of noise or bias from the data that might be due to carelessness, with perhaps the worst offenders being those who answered randomly. To accomplish this quickly, gross measures were searched for and used.

The replacement of data requires a careful balance between eliminating poor data due to gross noncompliance and preserving genuine variability. For example, keeping data from someone that randomly answered yes or no degrades the data. Removing data from an honest subject who happens to be an outlier will introduce bias into the data set. Thus, we sought to achieve a balanced rational for deciding which data to replace.

An effort was made to determine which participants answered questions randomly or without consideration. This analysis considered both the proportion of "yes" responses attributed to a participant as well as the individual's average score on check vignettes (described earlier). An analysis of the data suggested that individuals with a mean performance on the check vignettes

of less than 60% correct should be excluded from the study. This class of participants included those who always gave the same answer as well as those who answered as if they were choosing randomly. As indicated previously, data from this excluded class of participants were discarded and collected anew using MTurk, ensuring that the noncompliant individuals would not be participating further.

3.4.6.2 Data Validation

With the "cleansed" set of data in hand, the next order of business was to validate the data from the RT:CS using the RT:PS data. Toward this end, a chi-square test was applied to the analysis two by two contingency tables to compare the distribution of present/absent responses across data sets. The use of the Pearson chi-square statistic is based on the formula

$$\chi^2 = \sum \frac{(observed - \exp ected)^2}{\exp ected} \quad (1)$$

In the RT:PS, a representative set of 24 (out of the original 480) exemplars were used. For purposes of comparison, only the portion of the Crowdsourced Study data set corresponding to those 24 verbs was used in the analysis. The data from each set were split into 24 x 48 cells corresponding to each combination of exemplar and verb question. Each of these cells was compared across data sets. Thus, a two by two contingency table generated a chi-square value (with one degree of freedom) for each comparison. The sum of the chi square values for each cell in the matrix of exemplars and verb questions led to a chi-square test with 48*24 = 1152 df. The overall chi-square test on the entire matrix had a p-value of 0.4205, indicating there is not enough evidence to reject the hypothesis that the data sets are similar.

This was intended to be a preliminary analysis for the purpose of validating both the methodology and resulting data from the RT:CS. It appears that the data from the two studies are consistent with each other, and furthermore, that the replacement of some individuals' data has improved the overall quality of the data set. On this basis of confidence, we have released partial data from the RT:CS for use in system development and will use the remainder for evaluation in year 1.

3.4.6.3 Future Analyses

Two related areas of inquiry will be explored. One is to compare quantitatively the similarity of human and system data. The other is to characterize human and system data qualitatively. The reason these analyses are related is that in order to make meaningful quantitative comparisons we need to rely on the results of the qualitative analyses. So while a purely quantitative comparison between human and system performance can and will be made, considering comprehensively the outputs of both systems and the sum of squared error that characterizes the differences between those outputs, we will also use methods to identify groupings of perceptually similar stimuli and examine how systems and humans differ in their treatment of those groupings.

Candidate methods for characterizing the similarity space of a perceptual system derive from confusion matrix representations. Among these methods are dimensionality reduction techniques such as multidimensional scaling and cluster analysis. One way to visualize the difference between two perceptual mappings is to derive the high-to-low dimensional mapping for one space and then use that mapping to overlay percepts from the other space. Thus, it may be worthwhile to derive a similarity space from the human response data and then overlay system data into that space. Such an overlay could reveal interesting differences that could then be targeted for deeper analysis.

Another line of inquiry will seek to identify factors that affect system performance but are not likely related to visual intelligence. A likely approach to this would be to perform a multivariate analysis of variance (ANOVA) that examines the effects of production method (e.g., live-action vs. composited vs. animated video) as well as the four dimensions of variation used to produce the 16 variants associated with each exemplar.

4. Human Response Data: Description Task

4.1 Description Task: Vignette Descriptions Collected from Panel of Human Subjects

4.1.1 Rationale

The Description Task Panel Study (DT:PS) was intended to provide the system design teams with samples of short descriptions for video vignettes provided by trusted human reviewers. A "short description" was limited to 140 characters, sufficient for two or three short sentences. The panel members, after viewing a video, had to make choices as to what actions in the vignette to describe and which verbs to use for the description from the list of 48 verbs.

4.1.2 Participants

The ARL panel makeup was composed of five engineers. All panel members were familiar with the goals of the program and were considered expert reviewers. The assumption was that these expert reviewers would put forth the best effort for the human response data. These data were later augmented by crowdsourced data.

4.1.3 Stimuli

The vignettes used for the DT:PS were selected from the vignettes created for the RT:PS. Two engineers reviewed the vignettes at the exemplar level to see which vignettes were the best candidates for description. There were many vignettes that were determined to be not suitable because of the complexity and nature of the actions depicted.

Thus, 480 vignettes from the Recognition Task vignettes were selected for the Description Task. Half of these vignettes (240) were released to the system design principal investigators (PIs),

along with HR data, for training their systems. The remaining 240 vignettes will be used to evaluate the VI systems.

4.1.4 Method

The DT:PS collected description data from the expert panel. The panel was given instructions, presented on the GUI, each time a new vignette was presented. It was also expected that the panel would draw on their previous experience with the ARL Recognition Task studies. No additional training was provided. Definitions for the verbs were displayed as the mouse cursor scrolled over the verb. A panel member would watch a video then author a description, using the verbs shown as much as possible. When the description was complete, the panel member hit the submit button and the next vignette was displayed (figure 9).

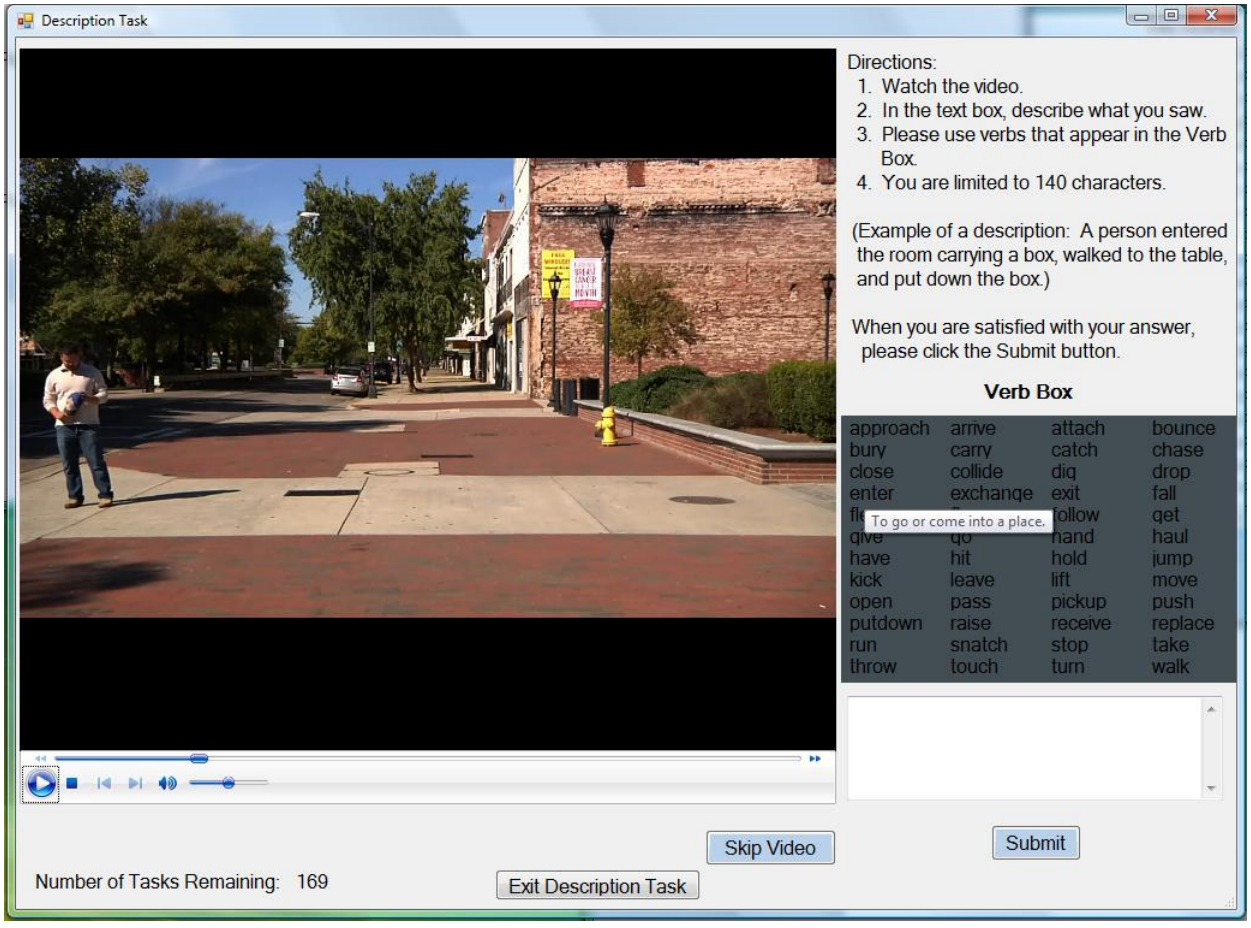


Figure 9. ARL Description Task GUI.

4.1.5 Resulting Data

A total of 480 vignettes were handpicked for the study. Each was displayed to 5 panel members to generate 5 total responses per vignette. This procedure resulted in 2,400 responses that were then divided equally between development and evaluation sets.

All panel members were given the option to use all 48 verbs to describe the scene. The reviewers were trusted to make the best verb selection based on their judgment. Other than human review of the data by test examiners, there would be no other quality control on the data for year 1.

4.1.6 Assessment

A comparison between the DT:PS data set and the descriptions collected via crowdsourcing (described in section 3.2) shows that these data have less variability than the crowdsourcing data set, hewing more closely to the year 1 verbs of interest and simple description grammar. The DT:PS data were deemed as more useful examples of acceptable natural-language descriptions, against which systems can be subjectively compared.

4.2 Description Task: Vignette Descriptions Collected via Crowdsourcing

4.2.1 Rationale

The Description Task Crowdsourced Study (DT:CS) effort was intended to provide more data to the system design community, with greater variability than was achievable in the ARL panel study for the DT:PS. As with the ARL panel, the short descriptions were limited to 140 characters. It was assumed that crowd sourcing the Description Task would produce variability that includes, to some extent, exceeding the vocabulary chosen in year 1, and that this variability would be useful in year 1 as an exploration of HRs.

In addition to producing more sample descriptions, the DT:CS was also viewed as an opportunity to explore verb classes that may be omitted from the list of 48 verbs of interest and of potential interest in future years. MTurk Workers were instructed to use the 48 verbs as much as possible but the use of other verbs was not prohibited. Straying from the 48 verbs seemed illustrative as long as the description was correct.

4.2.2 Participants

Through the MTurk site, access to the DT:CS HITs was restricted to MTurk Workers residing in the United States and who had a 95% accuracy rating. It is assumed that the subtleties in the English language would be better addressed by a population that regularly uses English. We also kept the requirement of a 95% accuracy rating from the RT:CS. Personal accuracy rating was highly prized by each MTurk Worker and they were very conscientious about maintaining the rating. This added motivation would help ensure that the MTurk Worker doing a Description HIT would strive for accuracy and reduce the probability of a Worker contributing spurious data. No identifying information was maintained on the MTurk Workers beyond the user IDs assigned for data storage in the ARL database.

4.2.3 Stimuli

The vignettes used for the DT:CS were selected from the vignettes created for the RT:PS and were exactly the same as those used for the ARL DT:PS. Two engineers reviewed the vignettes

at the exemplar level to identify vignettes to be used in the study. Vignettes were selected to avoid complexity and ambiguity in the nature of the actions depicted. 480 vignettes were selected in total for the DT:CS.

4.2.4 Method

A Description Task HIT presents each MTurk Worker with a vignette along with instructions which directed the Worker to watch the video and then author a brief description of what they had seen. Their responses were limited to 140 characters, which they typed in the text box provided (figure 10). A list of verb definitions was available via a separate link. It should be noted that, due to implementation issues, it was not possible to replicate the "mouse-over" access to definitions that were available to participants in the Panel Study, which is why a separate link was used. No additional training was provided.

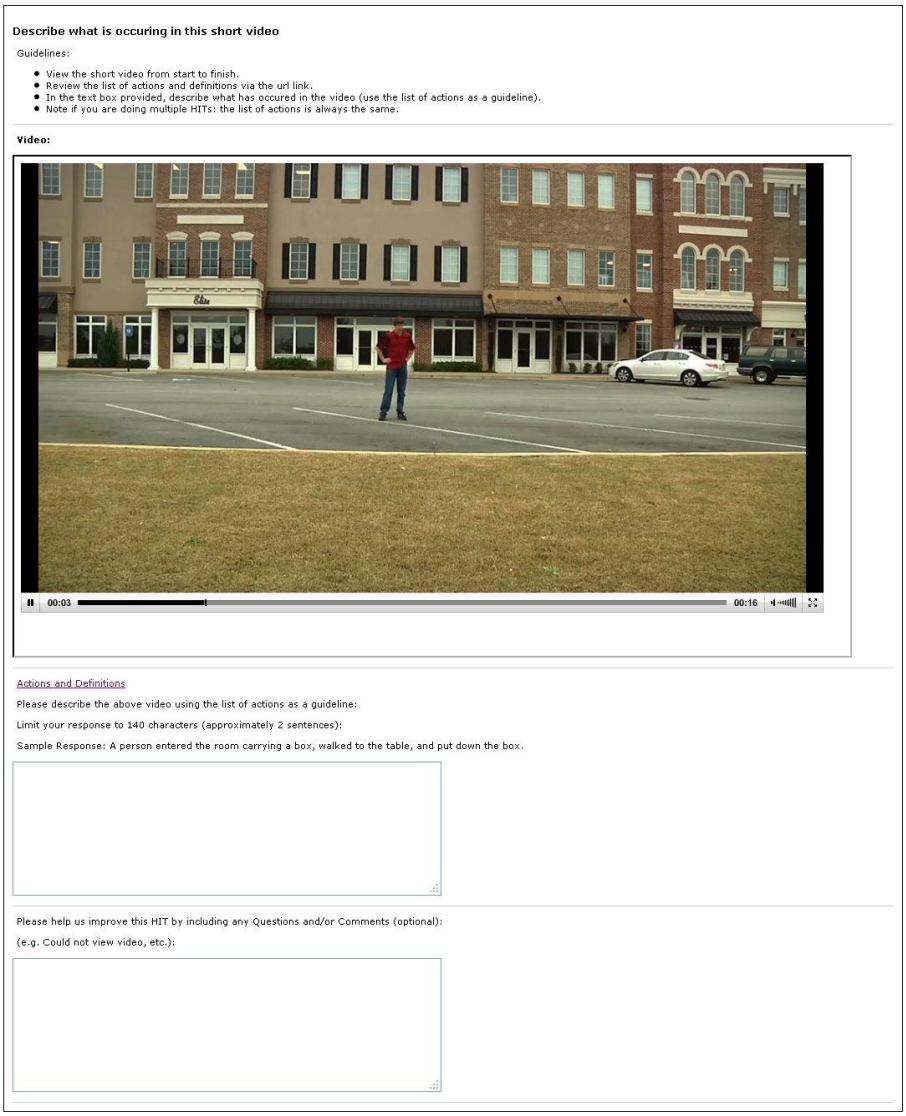


Figure 10. MTurk Description Task GUI.

Unlike the RT:CS study, each HIT was given 10 assignments, meaning that 10 different Workers could provide a response for each HIT video.

4.2.5 Resulting Data

A HIT was created for 480 hand-picked vignettes with each HIT soliciting responses from 10 different Workers. This setup resulted in 4,800 responses that were then divided equally between Development and Evaluation use.

Manual inspection of the data by test reviewers revealed that while MTurk Workers may not have used verbs from the 48 verb set, the Workers did provide accurate descriptions of the scenes. There was no quality control enforced for this task other than a review for blank or nonsensical responses.

Worker use of the verb definition page was not recorded. Participant-based variability in the use of this information may have contributed to response variation.

4.2.6 Assessment

The DT:CS data have not yet been formally analyzed. To assess the accuracy of the responses and generate a conclusion, a combination of automated and manual methods will be used to summarize the data collected in this study. For example, key-word checking for the 48 verbs including tenses and synonyms will provide a starting point for describing the data. These descriptions will be contrasted with the detection data profiles associated with the description vignettes.

Information retrieval methods will be applied to the evaluation of system data with respect to these data, with some human in the loop to treat ambiguous cases. It is expected that the representations used in system responses will span a range of abstraction and descriptiveness; therefore, the analysis of these data will necessitate some degree of subjective analysis.

4.3 Gap-Filling Task: Responses Collected via Crowdsourcing

4.3.1 Rationale

By collecting data through crowdsourcing, we would be able to provide performers with many more examples of how humans exercise their ability to reason across gaps in visual information. The Gap-filling Task: Crowdsource Study (GF:CS) tasked MTurk Workers to describe the most plausible action to use for a video that is missing a segment. Using vignettes from the Description and Evaluation sets, the video files were altered to display nothing but a black screen for a portion of the video. It was assumed crowdsourcing a description effort would lead to a great amount of variability in the data.

Another rationale for crowdsourcing this task was to collect data on verb usage. Even though the MTurk Workers were instructed to use the 48 verbs, the interface did not prohibit them from

using verbs other than the list of 48 verbs. As long as the descriptions were accurate, the verbs they chose to use may suggest expansions to the verbs of interest.

4.3.2 Participants

Through the MTurk site, access to this task was restricted to MTurk Workers residing in the United States and who had a 95% accuracy rating. As with the Recognition and Description Tasks, workers with a good familiarity with the English language were preferred. The 95% accuracy rating requirement was instituted to have a barrier of entry for Workers to filter out those that have traditionally provided low accuracy work for other HITs. In addition, personal accuracy rating is highly prized by MTurk Workers. As such, they are very conscientious about maintaining their rating and will strive to provide accurate data in order to keep their high accuracy rating. No identifying information was maintained on the MTurk Workers beyond the user IDs assigned for data storage in the ARL database.

4.3.3 Stimuli

The Gap-filling vignettes were created at ARL from the RT:PS vignettes. In order to evaluate human performance on prediction, interpolation, and postdiction, there were three types of gaps introduced in the vignettes. Prediction was examined by introducing a gap near the end of the video. In the case of interpolation, either a long or a short gap was introduced in the middle of the vignette. The different lengths allowed for evaluation of the impact of the length of the gap on performance. Postdiction was evaluated by placing a gap near the beginning of the video and asking the human reviewers to describe what may have happened during the gap.

Two ARL engineers reviewed the vignettes at the exemplar level to see which vignettes were the best candidates for gap-filling. A judgment was made on the individual vignettes as to whether the action in the vignette lent itself to the Gap-filling Task. A second judgment was made as to whether a selected vignette would be best used for prediction, interpolation, or postdiction. In all, 80 vignettes from the Recognition Task set were selected for the Gap-filling Task, with 20 designated as prediction videos, 20 as postdiction videos, and 40 as interpolation videos. From these 80 vignettes, 120 gap-filling vignettes were created. The increase in number was due to the creation of two sets of interpolation videos—40 interpolation videos were created by introducing a short gap into the original source video and an additional 40 were created by inserting a long gap into the same source videos. All 120 gap-filling videos were available as HITs on MTurk, and those HITs remained available until 10 responses were collected per video.

4.3.4 Method

As with the Description task, the short description was limited to 140 characters. An MTurk Worker would watch a vignette in which an artificial gap had been placed, and then imagine the most plausible action that could have occurred during the gap and author a description for the gap using the 48 verbs when possible.

The MTurk Workers had instructions presented through the GUI each time a new vignette was presented (figure 11). No additional training was provided. A list of verb definitions was made available in a separate link. Due to implementation issues, it was not possible to replicate the "mouse-over" access to definitions that were available to participants in the Panel Study, which is why a separate link was used.

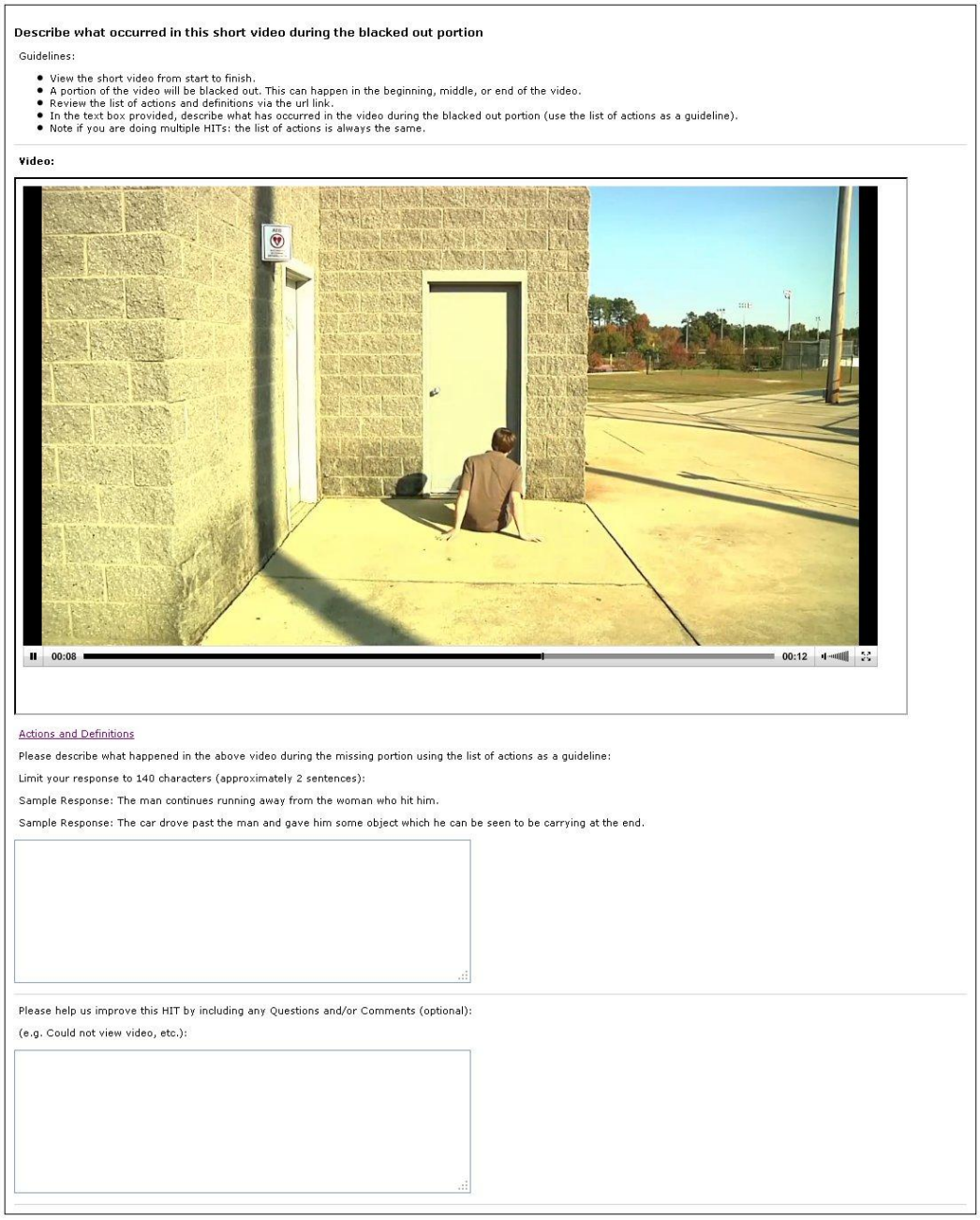


Figure 11. MTurk Gap-filling Task GUI.

All of the Gap-filling videos were available as HITs through MTurk. The HITs remained active until 10 responses had been collected per video.

4.3.5 Resulting Data

There were 80 hand-picked vignettes from which 120 vignettes were created. Each created vignette had 10 responses. This resulted in 1,200 responses that were then divided equally between development and evaluation purposes

The MTurk Workers may not have used a verb within the 48 verbs defined but overall the data accurately describes the scene. There was no quality control enforced for this task other than manual inspection for blank or nonsensical responses.

Whether the MTurk Workers took advantage of the definition link page was not recorded. This also may have been a factor in the variability in the responses.

4.3.6 Assessment

In future studies, a combination of automated and manual methods will be used to summarize the data collected in this study. For example, keyword checking for the 48 verbs including tenses and synonyms will provide a starting point for characterizing this data set.

Due to the variability in response representation, the comparison of system outputs to the human gap-filling data will likely involve a combination of objective and subjective methods. At a minimum, a comparison of verbs identified to occur in the gaps will be made.

5. Conclusion

As this study necessitated the acquisition of large data sets, crowdsourcing techniques were the primary focus for data collection. The basic cost versus benefit dichotomy was identified as time requirement and convenience versus accuracy. Several techniques were used to minimize accuracy concerns, including the gold standard baseline questions and the MTurk worker qualification requirements.

Several conclusions can be drawn from the procedure documented in this report. Perhaps the most significant is that precautions must be taken to combat against malicious users. Analysis of the collected data from the crowdsourced recognition task showed approximately 8% of the data to be fraudulent. In addition, proper incentive must be applied. HITs with higher compensation were completed at a noticeably faster rate than HITs with lower compensation. Communication with the MTurk worker pool was vital for establishing a fair reward value.

Once established, MTurk proved to be an efficient source of data. Estimates to complete the data collection from ARL employees were multiple months in length. Using MTurk, the data collection was completed over several collection periods totaling fewer than 30 days. While more in depth analysis is required to establish accuracy, initial review of the data shows an acceptable level of accuracy to justify this method of crowdsourcing.

List of Symbols, Abbreviations, and Acronyms

ANOVA	analysis of variance
API	application programming interface
ARL	U.S. Army Research Laboratory
BEP	Broad Evaluation Plan
DT:CS	Description Task Crowdsourced Study
DT:PS	Description Task Panel Study
GF:CS	Gap-filling Task: Crowdsource Study
GUI	graphical user interface
HITs	Human Intelligence Tasks
HR	human response
HTML	hypertext markup language
MTurk	Amazon Mechanical Turk
PI	principal investigator
RT: CS	Recognition Task: Crowdsourced Study
RT: PS	Recognition Task: Panel Study
RT: RT	Recognition Task: Round Table
S3	Amazon Simple Storage Service
SVPA	Single Verb Present/Absent
XHTML	extensible hypertext markup language
XML	extensible markup language

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) DIRECTOR US ARMY RESEARCH LAB IMAL HRA

1 (PDF) DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL

1 (PDF) CHIEF US ARMY RESEARCH LAB
1 (HC) RDRL CII 2800 POWDER MILL RD ADELPHI MD 20783-1197

1 (PDF) STUART YOUNG US ARMY RESEARCH LAB RDRL CII A

1 (PDF) ROBERT WINKLER US ARMY RESEARCH LAB RDRL CII B

1 (PDF) LAUREL SADLER US ARMY RESEARCH LAB RDRL CII B

1 (PDF) NICHOLAS FUNG US ARMY RESEARCH LAB
1 (HC) RDRL CII A 2800 POWDER MILL RD ADELPHI MD 20783-1197